数据挖掘在铁路客流分析预测中的应用①

Application of Data Mining in the Analysis and Forecasting of Railway Passenger Flow

郑 丹 王 耀 (华东交通大学 职业技术学院 江西 南昌 330002)

摘 要: 应用数据仓库和数据挖掘技术，以铁路客票发售和预订系统为研究主体，将有效的数据挖掘技术应用于铁路客流分析，采用神经网络思想，建立了一个基于 BP 神经网络模型的客流分析预测模型，为客运部门合理安排运能、科学组织管理提供了准确的决策信息和先进的预测手段。

关键词: 数据仓库 数据挖掘 BP 神经网络 客流 铁路客票发售和预订系统

中国铁路客票发售和预订系统的建设和联网售票的实现，极大地方便了旅客购票，有力推动了客运的改革，同时也积累了大量客票发售的生产业务数据。这些数据规模庞大，蕴涵着丰富的决策支持信息。如果能开发这些宝贵的信息资源，为组织优化客流、运输方案制订，铁路客流计划的编制和客流统计分析、预测、以及辅助决策服务，将会带来巨大的经济效益。

1 引言

1.1 文章安排

本文第 2 节介绍客流分析预测系统的实施过程。第 3 节给出详细的预测模型的建立过程及实验预测结果。

1.1.1 基本介绍

本文以铁路客票发售和预订系统为研究主体，采用神经网络思想，从大量售票数据中挖掘出隐藏于其中的有用信息,提出了一个基于 BP 神经网络模型的客流分析预测模型，并利用此模型对 2007 年“五一”黄金周的客流进行预测，通过预测值与实际值的比较对客流情况作出评估，最终的挖掘与分析信息可反馈给客流组织有关的部门，根据这些信息，决策者可以调整和改进客流计划，管理者及时做好客流的组织，分流和优化工作，为旅客提供更优质的服务[1]。

2 客流分析预测系统的实施过程

2.1 数据源构成

由于客票系统的数据库和复制服务器均采用的是 SYBASE 的产品，而本文构建数据仓库将采用 Microsoft 公司的 SQL Server 2000，所以存在异构数据源的转换问题。技术上采用 PowerBuilder 与 Sybase 数据库连接，将所需要的基础表导入至 SQL Server 中。

2.2 创建数据集市

对有些企业，尤其是中小型企业更倾向于在不影响信息系统的基础上，建立小型的数据集市，进行有针对性的主题挖掘。本文通过创建数据集市，将相关数据抽取到数据集市中，再将数据集市中多个表的数据抽取到一个指定的关系表中然后在此基础上利用 OLAP 工具具建立一个多维分析模型(立方体)，再进行数据挖掘，实现铁路客流的分析预测。

3 客流分析预测模型的建立

3.1 样本数据

由于前期已经建立基于本主题的数据集市，从中可以提取出相关数据进行数据挖掘。在数据集市中包含了各大站、区段和干线的售票量，客票收入，运输收入、旅客上车人数等各种分类别统计的运量信息。

① 收稿时间:2009-02-07

神经网络学习前的数据处理对网络有至关重要的影响，它可以影响到网络的学习速度以及网络的精度等。通常获取的数据样本不是都能直接用于网络的训练，而需要对原始数据进行一定的处理。

3.2 BP 神经网络中的数据预处理

数据变换是数据预处理的一个重要部分。数据变换就是将数据转换成适合于挖掘的形式，即是将特征向量数据按比例缩放，使之落入一个小的特定区间。常用的处理方法是归一化处理。在本次的网络模型计算中，输入样本和检验样本的数据均统一量化为(0，1)之间的实数。归一化的方法有很多种，这里采用如下公式[2]:

$$\hat{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

利用公式(1)在 MATLAB 中实现向量的归一化，这里将样本归一化到区间[0,1]。

3.3 客流预测模型的建立

本文采用数据挖掘中的神经网络方法进行客流数据建模。铁路客流量变化具有复杂性、随机性和周期性的特征。铁路客流的周期性一般是以年为周期，每年的春运、暑期、“五一”，“十一”黄金周呈季节性变化。

该次实验以南昌火车站 2005 至 2007 三年“五一”黄金周数据为例，通过近两年的数据，预测 2007 年“五一”黄金周的旅客客流量。通过实验及实验结果，希望能预测到 2007 年“五一”节期间的客流量，为客流组织，运输方案的制定提供参考。

这里分别取售票量，客票收入，运输收入作为影响因子、以旅客上车人数作为输出因子，即网络的输出。

本文采 BP 神经网络建立模型,根据上面确定的网络输入和输出因子，利用 2005-4-28～2006-5-7 年共 20 条历史统计数据作为网络的训练样本，2007-4-28～2007-5-7 即“五一”节期间的历史统计数据作为网络的外推测试样本[3]。具体的样本数据如表 1 所示(篇幅所限，部分数据省略)。

3.4 预测指标的选择

在客流预测中，预侧指标为高峰时期的旅客上车人数。

3.5 BP 人工神经网络隐含层的节点选择

基于 BP 算法的神经网络中各层节点数目的选择对于网络的性能影响很大，过多的隐含层节点对网络的概括推理能力产生不利影响，即影响网络对于新输入的适应性。而过少的隐含层节点数目会影响网络学习的精确度并且使网络学习出现局部极小的情况增多，所以层内部节点数需要进行恰当的选择。

表 1　样本数据(输入样本、目标样本)

样本日期	售票量	客票收入	运输收入	旅客上车人数
2005-4-28	39337	2671330	2822110	26937
……	……	……	……	……
2006-5-7	37298	2117053	2241152	47668
2007-4-28	47122	3690184	3897888	48529
……	……	……	……	……
2007-5-7	32304	2074047	2208837	49916

常用的解决办法就是使隐含层单元数目可变。一种是开始放入足够的隐含单元，然后把学习后那些不起作用的隐含层单元逐步去掉，一直减少到不可收缩为止。另一种是开始放入比较少的隐含层单元，学习一定次数后，还不成功就要增加隐含单元个数，一直达到比较合理的隐含单元数目为止。

本实验网络的输入层神经元个数为 3，根据上面所讲的隐含层的设计，以及考虑本实验实际情况，解决该问题的网络的隐层神经元个数应该在 6~12 之间。因此,设计一个隐含层神经元数目可变的 BP 网络，通过误差对比，确定最佳的隐含层神经元个数，并检验隐含层神经元个数对网络性能的影响[4]。

网络的设计及训练代码如下。

```
s=6:12;
res=1:7;
for i=1:7
net=newff(minmax(P_train),[s(i),1],{'tansig',
'logsig'},'trainlm');
    net.trainParam.epochs=1000;
    net.trainParam.goal=0.001;
    net=train(net,P_train,T_train);
    y=sim(net,P_train);
    error=y-T_train;
    res(i)=norm(error)
end
```

代码的运行结果如表 2 所示。

表 2　网络训练误差

神经元个数	6	7	8	9
网络误差	0.1181	0.1134	0.1287	0.1250
神经元个数	10	11	12	
网络误差	0.1220	0.1299	0.1281	

表 2 表明，在经过 1000 次训练后，隐含层神经元为 7 的 BP 网络对函数的逼近效果最好，因为它的误差最小。

至此，确定了本实验的最终 BP 网络结构，如表 3 所示。

表 3　网络最终结构

网络结构	隐含层神经元
3×7×1	7 个
训练函数	网络误差
Trainlm	0.1134

3.6 BP 网络训练与测试

我们利用表 1 中的数据进行训练。训练后的网络才有可能满足实际应用的要求。

变量 P_train,T_train 分别表示网络的输入向量和目标向量，它们是从表 1 中得出的。训练结果为

TRAINLM, Epoch 0/1000, MSE 0.0970386/0.001, Gradient 0.811959/1e-010

TRAINLM,Epoch25/1000,MSE 0.00324957/0.001, Gradient 0.0260721/1e-010

TRAINLM,Epoch50/1000,MSE 0.00310333/0.001, Gradient 0.0110766/1e-010

TRAINLM,Epoch75/1000,MSE 0.00220873/0.001, Gradient 0.0036355/1e-010

TRAINLM,Epoch100/1000,MSE 0.001898-54/0.001, Gradient 0.00503494/1e-010

TRAINLM,Epoch116/1000,MSE 0.000970-053/0.001, Gradient 0.0578464/1e-010

TRAINLM, Performance goal met.

可见经过 116 次训练后，网络的目标误差达到要求，如图 1 所示。

网络训练结束后，还必须利用另外一组客流数据对其进行测试。

测试代码为：

y=sim(net,P_test);

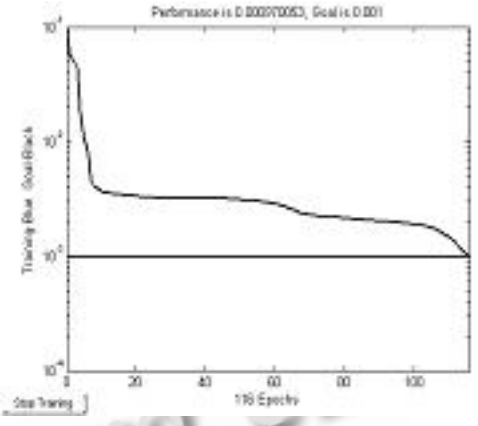

图 1　训练结果

输出结果经过反归一化处理后得到预测的 2007 年“五一”黄金周时期的客流量，和实际的旅客上车人数相比较可得到网络的预测误差，如表 4 所示。

表 4　预测结果对照

模型生成数据	实际数据	相对误差(%)
48584	48529	0.113334295
47250	49599	-4.73598258
54200	56124	-3.428123441
69858	66832	4.527771128
……	……	……

由表 4 可见，网络的预报误差比较小，都在 5% 左右。通过表的显示说明基于 MATLAB 神经网络工具箱的 BP 网络对铁路客流分析不仅在技术上是可行的，且结果是可靠的。

参考文献

1 徐薇,黄厚宽.基于时空数据挖掘的铁路客流预测方法.北京交通大学学报, 2004,28(5):16－19.

2 Zhao Y, Nan J, Cui FY. Water quality forecast through application of BP neural network at Yuqiao reservoir. Journal of Zhejiang University SCIENCE A, 2007, 8:1482－1487.

3 Zhu XL, Wu BD,Wu XX. Application of RBF Neural Network in Optimizing Machining Parameters. Journal of Shanghai University(English Edition), 2004, 8:106－110.

4 葛哲学,孙志强.神经网络理论与 MATLABR2007 实现.北京:电子工业出版社, 2007.